DOCUMENT RESUME

ED 236 191 TM 830 700

AUTHOR Tindal, Gerald; And Others
TITLE The Technical Adequacy of a Basal Reading Series Mastery Test.
INSTITUTION Minnesota Univ., Minneapolis. Inst. for Research on Learning Disabilities.
SPONS AGENCY Office of Special Education and Rehabilitative Services (ED), Washington, DC.
REPORT NO IRLD-RR-113
PUB DATE Apr 83
CONTRACT 300-80-0622
NOTE 43p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Basal Reading; Criterion Referenced Tests; Grade 6; Intermediate Grades; *Mastery Tests; Measurement Techniques; Reading Research; *Reading Tests; *Test Reliability; *Test Validity
IDENTIFIERS *Houghton Mifflin Reading Series; SRA Diagnostic Reading Tests; Word Reading Test

ABSTRACT

The purposes of this study were to examine the reliability and validity of a basal reading series mastery test, and to explore the appropriateness and usefulness of two strategies for investigating the reliability and validity of criterion-referenced tests. Subjects were 47 sixth graders, who were tested on the SRA Reading Achievement Test, the Houghton-Mifflin End-of-level 11 Basic Reading Test (BRT), and the Word Reading Test. A subgroup of 20 children was tested a second time on the BRT. Traditional psychometric correlational analyses as well as specific strategies for examining the adequacy of criterion-referenced tests were applied to the data to investigate the following dimensions of the technical adequacy of the BRT: (1) consistency of student performance across two administrations of the BRT, and (2) criterion validity of the BRT scores with respect to two other measures of reading proficiency and criterion validity of the BRT mastery/nonmastery decisions with respect to pre/post instructional status. Results indicated that the reliability and validity of the BRT was less than adequate, and that both strategies for investigating the adequacy of a criterion-referenced test were useful and provided complementary information. Implications for the development and use of criterion-referenced instruments are discussed. (Author)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED236191

University of Minnesota

Research Report No. 113

THE TECHNICAL ADEQUACY OF A BASAL READING SERIES MASTERY TEST

Gerald Tindal, Mark Shinn, Lynn Fuchs, Douglas Fuchs,
Stanley Deno, and Gary Germann

SCOPE OF INTEREST NOTICE
The ERIC Facility has assigned this document for processing to: TM EC
In our judgement, this document is also of interest to the clearinghouses noted to the right. Indexing should reflect their special points of view. CS

iRLD

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
J. Ysseldyke
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Institute for Research on Learning Disabilities

TM830700

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☒ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

IRLD

Director: James E. Ysseldyke

The Institute for Research on Learning Disabilities is supported by a contract (300-80-0622) with the Office of Special Education, Department of Education, through Title VI-G of Public Law 91-230. Institute investigators are conducting research on the assessment/decision-making/intervention process as it relates to learning disabled students.

During 1980-1983, Institute research focuses on four major areas:

- Referral
- Identification/Classification
- Intervention Planning and Progress Evaluation
- Outcome Evaluation

Additional information on the Institute's research objectives and activities may be obtained by writing to the Editor at the Institute (see Publications list for address).

The research reported herein was conducted under government sponsorship. Contractors are encouraged to express freely their professional judgment in the conduct of the project. Points of view or opinions stated do not, therefore, necessarily represent the official position of the Office of Special Education.

Research Report No. 113

THE TECHNICAL ADEQUACY OF A BASAL READING SERIES MASTERY TEST

Gerald Tindal, Mark Shinn, Lynn Fuchs, Douglas Fuchs,
Stanley Deno, and Gary Germann
Institute for Research on Learning Disabilities
University of Minnesota

April, 1983

Abstract

The purposes of this study were to (a) examine the reliability and validity of a basal reading series mastery test, and (b) explore the appropriateness and usefulness of two strategies for investigating the reliability and validity of criterion-referenced tests. Subjects were 47 sixth graders, who were tested on the SRA Reading Achievement Test, the Houghton-Mifflin End-of-level 11 Basic Reading Test (BRT), and the Word Reading Test. A subgroup of 20 children was tested a second time on the BRT. Traditional psychometric correlational analyses as well as specific strategies for examining the adequacy of criterion-referenced tests were applied to the data to investigate the following dimensions of the technical adequacy of the BRT: (a) consistency of student performance across two administrations of the BRT, and (b) criterion validity of the BRT scores with respect to two other measures of reading proficiency and criterion validity of the BRT mastery/nonmastery decisions with respect to pre/post instructional status. Results indicated that the reliability and validity of the BRT was less than adequate, and that both strategies for investigating the adequacy of a criterion-referenced test were useful and provided complimentary information. Implications for the development and use of criterion-referenced instruments are discussed.

The Technical Adequacy of a Basal Reading Series Mastery Test

With the growing demand for accountability in the schools, the focus on educational tests has expanded. Norm-referenced achievement testing, the traditional measurement format, is the predominant measurement strategy for evaluating and documenting program effects. Concurrent with its frequent use, however, is growing recognition that norm-referenced measurement may be inadequate for its intended purposes: It has poor content validity with respect to classroom curricula, and it fails to indicate the extent to which individuals or groups have mastered specific educational objectives (Skager, 1971).

As an alternative to traditional educational measurement, criterion-referenced (CR) testing has received greater attention in the past two decades by measurement theorists, test developers, and school personnel. As conceptualized by Glaser and Nitko (1971), the CR test is a sample of items yielding information that is interpretable directly with respect both to a well-defined domain of tasks and to specified performance standards. This definition reflects three characteristics that frequently are employed in the literature to describe CR measurement: (a) definition of a well-specified content domain (Baker, 1974; Hambleton & Novick, 1973; Millman, 1974), (b) delineation of valid performance criteria (Hambleton, 1980), and (c) development of procedures for generating appropriate samples of tests (Goodstein, 1982; Hambleton, Swaminathan, Algina, & Coulson, 1978; Popham, 1980). All three components stress the edumetric and psychometric properties of CR tests.

Nevertheless, the focus both in publishing houses and in the

schools has been more utilitarian. With the recognition that CR tests provide relevant data for describing student progress with respect to specific learning objectives, their use has proliferated. Test developers have marketed CR instruments along with objective banks; commercial curriculum writers have published CR tools for assessing mastery within their series; school districts have created their own CR tests; and teachers have developed such instruments to fit individual learning objectives. Unfortunately, there has been a lack of concommitant investigation of the reliability and validity of these tests.

Therefore, although two measurement formats currently are available and used in educational settings, neither is adequate for evaluating the effects of instructional programs. While norm-referenced tests frequently demonstrate several strong psychometric characteristics, they lack content validity and utility. Alternately, CR instruments are isomorphic with respect to classroom curricula and, as such, appear very useful; however, there is little evidence that such measurement is accurate or meaningful.

The current study addressed part of this dilemma by beginning the task of investigating the reliability and validity of available CR tests. Traditional ways of assessing such adequacy, however, have been criticized as largely inappropriate for CR instruments (Popham & Husek, 1969). Hambleton and Novick (1973) reasoned that, because one of the purposes of a CR test is to identify mastery within a domain, test variance typically is small. Homogeneous distributions of test scores are centered at the low and high ends of the measurement scale,

respectively representing pre and post-instruction performance (Hambleton & Novick, 1973). When the variance of test scores is restricted in this way, correlational estimates of reliability and validity tend to be low. In response to this problem, alternative analyses for investigating the adequacy of CR tests have been developed (Berk, 1980); in contrast to the correlation statistic, these analyses rely minimally on the notion that inter-individual variability is necessary (Carver, 1970; Hambleton & Novick, 1973; Huynh, 1976; Subkoviak, 1975).

Despite the development of such analyses, it appears that developers of commercial CR instruments, if they address technical adequacy at all, still rely predominantly on traditional psychometric correlational analyses. Inspection of eight commercial criterion-referenced instruments and four basal mastery tests revealed that (a) only one-third of the test manuals addressed reliability and validity at all, and (b) only traditional analyses were employed in the investigations of the instruments' technical adequacy (see Table 1).

Insert Table 1 about here

In the present study, both traditional correlational statistics and alternative CR approaches were employed to examine the adequacy of one CR instrument developed and published by a reading series company. The purpose of this study was twofold. First, the investigation was designed to contrast results based on the traditional and alternative approaches to studying the technical adequacy of CR instruments. Such

a contrast should shed light on the appropriateness and potential usefulness of each strategy. The second purpose was to describe the reliability and validity of the specific CR measure examined. Despite widespread use of this test, there are few, if any, reports concerning its adequacy.[1] The investigation of the test's reliability and validity should provide information of interest not only to consumers of this measure but also to users of other CR tests for which technical data also are still unavailable.

Method

Subjects

Subjects were 47 students (20 M, 27 F) from two sixth grade classes. Each class represented a school district within a rural midwestern educational cooperative. The students' mean reading percentile rank was 51.48 (SD = 18.11) as measured on the Science Research Associates (SRA) Reading Achievement Test.

Measures

Three measures of reading performance were used in the study: a basal series criterion-referenced test, a global norm-referenced test, and a curriculum-based word reading test.

Criterion-referenced test. Three scales of the End-of-level 11 Basic Reading Test (BRT; Brzeinski & Schoephoerster, 1974) of the Houghton-Mifflin basal reading series were employed as measures. Each of the three scales, Decoding Skills, Comprehension Skills, and Reference/Study Skills is comprised of several subtests. Table 2 lists the subtests constituting each scale and provides brief descriptions of tasks the examinee is required to do within each

subtest. This BRT is designed as a criterion-referenced test, with items per subtest ranging from 6 to 12 and with mastery-nonmastery cutoff scores established at 83% to 85% correct responses.

Insert Table 2 about here

Norm-referenced test. The Science Research Associates (SRA) Reading Achievement Test (Naslund, Thorpe, & Lefever, 1978) is comprised of two subtests: vocabulary and comprehension. In the vocabulary section, examinees are required to select, from four alternatives, a synonym for an underlined word in a sentence. In the comprehension section, examinees read 200-300 word passages and answer questions in a multiple choice format. Total test score is based on a linear combination of the two subtests. Internal consistency reliability was reported at .88 (Salvia & Ysseldyke, 1981).

Curriculum-based word reading test. The Word Reading Test (Deno, Mirkin, & Chiang, 1982) requires children to read aloud passages and isolated word lists and is scored in terms of average numbers of words correct and incorrect over two alternate forms of the Isolated Word Reading and Passage Reading scales. The 200-word passages are drawn randomly from a student's grade-appropriate level basal reading book; the 150-word lists sample words randomly from basals, with 60% of words drawn from the student's grade-appropriate level and 40% sampled equally from all previous levels. For the passage and isolated Word Reading Test, test-retest and alternate form reliabilities were at least .90 (Fuchs, Deno, & Marston, in press; Fuchs, Wesson, Tindal,

Mirkin, & Deno, 1981).

Procedure

All students were tested in groups, by a school psychologist for the SRA Reading Achievement Test, and by their classroom teachers for the BRT. The Word Reading Test was administered individually by trained aides. Standardized administration procedures were followed on all tests. Testing time ranged from 60 to 90 minutes for the SRA test, 60 to 90 minutes for the BRT, and five to six minutes for the Word Reading Test. All testing was completed within a two-week period.

To assess test-retest reliability questions, a subgroup of 20 students (11 M, 9 F) was administered the measures in the following order: BRT, SRA Reading Achievement Test, Word Reading Test, and BRT again. For the remaining, 27 students, each measure was given one time, with the order of administration random.

Data Analysis

Consistency of performance on two administrations of the same test. Consistency of students' performance on the BRT was assessed in three ways. In all three analyses, the students who had been tested twice on the BRT (N=20) were the subjects. First, traditional test-retest reliability was determined by correlating scores from the two administrations of the BRT. The other two analysis strategies were designed specifically for criterion-referenced measures (see Millman, 1974). In the first of these, consistency of students' subtest scores was determined by (a) computing individuals' percentage correct score on each subtest for each administration of the BRT, (b) calculating

for each individual his/her difference score across the two administrations of each subtest, and (c) determining the percentages of examinees having each possible difference score on each subtest. In the second strategy, consistency of mastery-nonmastery decisions on subtests was determined by dividing the difference between observed and chance proportions of agreements in decisions by the maximum value that difference could assume. (The chance proportion of agreements was computed by multiplying and then summing the marginal proportions of the same decision categories for the two administrations, as done in a chi-square test of association.)

Criterion validity. The criterion validity of the BRT was determined in two ways, employing the entire group of subjects (N=47). The traditional psychometric strategy of correlating scores on the measure of interest (BRT) with criterion measures was used. The SRA Reading Achievement Test and the Word Reading Test were employed as the criterion measures. Additionally, chi-square statistical tests were applied to contingency tables wherein mastery-nonmastery represented one dimension of each table and pre-post instructional status represented the other dimension. Percentages of misclassifications supplemented the chi-square tests.

Results

Table 3 is a display of students' mean scores and standard deviations on each subtest of the BRT, on each subscale and the total of the SRA Reading Achievement Test, and on the isolated word reading and passage reading scales of the Word Reading Test.

Insert Table 3 about here

Consistency of Performance on Administrations of the Same Test

Test-retest reliability correlations on subtests of the BRT are displayed in Table 4. For the decoding subtests, correlations were low, ranging from .20 to .42; for the comprehension subtests, correlations were low to moderate, ranging from .03 to .83; and for the study/reference skills subtests, correlations were high, ranging between .86 and .94.

Insert Table 4 about here

The second analysis of the consistency of performance involved calculating the percentages of examinees who had different percentage correct scores across the two administrations of the BRT. Figures 1-4 are graphic displays of the percentages of examinees displaying various difference scores on each subtest of the BRT; Table 5 summarizes the information illustrated on the graphs. The range of difference scores on the subtests fell between 0 and 83%. The percentage of examinees with 0% difference scores on two administrations ranged from 22 on an information appraising subtest to 85 on the word attack subtest. Across the decoding subtests, the mean percentage of examinees with 0% differences scores was 65 (SD = 28.28); across the comprehension subtests, the mean percentage was 57.20 (SD = 14.96); across the study/reference skills subtests, the

mean percentage was 51.25 (SD = 18.76); and across all the subtests, the mean percentage was 55.07 (SD = 17.92).

Insert Figures 1-4 and Table 5 about here

The third analysis of the consistency of performance addressed consistency of mastery-nonmastery decisions across the two administrations of the BRT. Table 6 is a display of the uncorrected and corrected proportions of examinees placed into the same decision category on the two administrations. On the decoding subtests, the corrected proportions are low, with the proportion of agreement on the Word Attack subtest 6% lower than chance and the proportion of agreement on the Pronunciation subtest only 18% greater than chance. On the comprehension subtests, the proportions of agreement were quite variable, ranging from 15% lower than chance to 88% greater than chance. On the study/reference skills subtests, proportions of agreement were moderate to high, ranging from 51% to 78% greater than chance.

Insert Table 6 about here

Criterion Validity

Correlational analyses were conducted between the BRT subtests and two criterion measures, the SRA Reading Achievement Test and the Word Reading Test. Correlations between the BRT subtest and the SRA subscale and total test scores are displayed in Table 7. They ranged

from .35 to .73 when SRA vocabulary subscale scores were involved, from .19 to .70 when SRA comprehension subscale scores were employed, and from .26 to .75 when SRA total scores were used. The average correlation for BRT decoding subtests was .41 (SD = .02); for BRT comprehension subtests, the average correlation was .52 (SD = .21), and for BRT study/reference skills subtests, it was .57 (SD = .07).

Insert Table 7 about here

Correlations between the BRT subtests and the Word Reading Test subscale scores are displayed in Table 8. They ranged from .27 to .57 when isolated word reading scores were involved, and from .31 to .68 when passage reading scores were employed. The mean correlation for the BRT decoding subtests was .34 (SD = .08); for the BRT comprehension subtests, the mean correlation was .47 (SD = .13), and for the BRT study/reference skills subtests, it was .56 (SD = .06).

Insert Table 8 about here

Criterion validity also was examined by inspecting the relation between mastery-nonmastery decisions on the BRT and actual pre-post instructional status. Relevant chi-square values, p-values, and percentages of misclassified students are displayed in Table 9. Across the decoding subtests of the BRT, the average percentage of misclassified students was 40.50 (SD = 3.54); across the comprehension subtests, the average percentage was 39.00 (SD = 4.58), across the

study/reference skills subtests, it was 23.33 (SD = 8.51), and across all the subtests, it was 33.50 (SD = 9.99).

Insert Table 9 about here

Discussion

The purpose of the current study was twofold. First, the study was designed to describe the reliability and validity of a criterion-referenced mastery test of a basal reading series. Second, by examining this reliability and validity, both with traditional correlational analyses and with alternative strategies developed specifically for criterion-referenced instruments, this investigation sought to contrast results and assess the appropriateness and potential usefulness of each strategy.

With respect to its first purpose, the study examined two aspects of the technical adequacy of the Houghton-Mifflin End-of-level 11 Basic Reading Test: the consistency of students' performance on two administrations of the test, and the criterion validity of the test. On both of these indices, the Houghton-Mifflin BRT appeared inadequate.

Test-retest reliability coefficients indicated that, when the BRT was administered twice within a short time interval, students' performance was very inconsistent on the decoding subtests; none of the correlations obtained for the decoding subtests even fell within the acceptable range for making group decisions (Salvia & Ysseldyke, 1981). On the comprehension subtests, correlations were poor to fair,

with the correlation for only one subtest, Meaning Acquisition, falling into the acceptable range for group decision making and with none of the correlations high enough for making decisions about individual students. On the study/reference skills subtests, however, student performance was more consistent, with all correlations .86 or better.

Results of this traditional correlational analysis of consistency of student performance across tests were corroborated with the criterion-referenced strategy of examining the proportions of examinees consistently classified into the same decision category. As with the correlational analyses, on the decoding subtests the proportions were low, at an average of only 6% better than chance agreement. On the comprehension subtests, proportions were low to moderate, with 57% greater than chance agreement on Literal Comprehension, 15% less than chance agreement on Interpretative Thinking, and a mean 62.33% greater than chance agreement on Meaning Acquisition. On the study/reference skills subtests, proportions were moderate to high with an average 66.25% greater than chance agreement.

When inspecting the consistency of test scores displayed in Figures 1-4, and in Table 5, the percentages of examinees scoring the same across two administrations of the BRT appear variable. There was no identifiable pattern within BRT scales; the average percentage of subjects scoring the same across all the subtests was 55. Given the fact that there are only 6 to 12 items per subtest and given a mastery criterion of 83% to 85% per subtest, a difference of one or two items correct in an administration of the BRT subtest can result in

different mastery decisions. Thus, an average of 55% of subjects scoring the same on two BRT administrations appears to be lower than desirable.

The results of the three analyses indicate that the consistency of student performance on the BRT is less than adequate and that educators should exercise caution as they attempt, on the basis of one administration of the BRT, to formulate decisions concerning whether individual students should progress to more difficult instructional material. While the study/reference skills subtests may be adequate as a data base for making such decisions, the decoding and comprehension subtests, which teachers may consider more critical for formulating decisions about reading proficiency, were unreliable.

The criterion validity of the BRT also was examined. The traditional correlational analyses indicated that the criterion validity of the BRT with respect to the SRA Reading Achievement Test and the Word Reading Test was poor to fair, with correlations falling between .19 and .73. Correlations on the Interpretive Thinking comprehension subtest were the lowest. Statistics for the decoding subtests also were relatively low, whereas the figures for the remaining comprehension and study/reference skills subtests were somewhat higher. Correlations among measures of reading proficiency frequently have been reported at high levels (Fuchs, Deno, & Marston, in press; Fuchs, Fuchs, & Deno, 1982). This indicates that the figures for the BRT are comparatively low and that performance on the BRT is a relatively poor predictor of concurrent performance on other measures of reading proficiency.

The criterion validity of the BRT also was investigated with the criterion-referenced strategy of examining the relation between the mastery-nonmastery classification on the BRT and actual pre-post instructional status. Relatively high percentages of misclassifications (15% to 43%) were found, suggesting limited utility of the BRT for classifying students into groups for instruction within the basal reader for which the BRT was designed.

Consequently, the current study casts doubt on the reliability and validity of the Houghton-Mifflin End-of-level 11 Basic Reading Test, and suggests that educators use this test with caution. Educational tests are designed to sample an individual's behavior, as a basis for drawing generalizations concerning his/her functioning and for making instructional decisions. When tests sample behavior in meaningful (valid) and accurate (reliable) ways, they are useful for such purposes. Although criterion-referenced tests may possess high content and face validity, their meaningfulness and accuracy remain empirical questions, an issue frequently ignored by criterion-referenced test developers. By investigating the reliability and validity of one criterion-referenced test, the present study (a) documents the notion that content validity is a necessary, but insufficient aspect of criterion-referenced test adequacy, and (b) underscores the importance of investigating the reliability and validity of criterion-referenced tests as they are developed.

The second purpose of this study was to compare the appropriateness and usefulness of traditional analyses with strategies developed specifically for criterion-referenced tests. Findings

discussed above suggest that the two types of analyses tend to corroborate and enhance each other, providing complimentary information. It appears that both strategies may be appropriate and necessary for investigating and describing the reliability and validity of criterion-referenced tests.

References

Baker, E. L. 3eyond objectives: Domain-referenced tests for evaluation and instructional improvement. Educational Technology, 1974, 14, 10-16.

Berk, R. A. A consumer's' guide to criterion-referenced test reliability. Journal of Educational Measurement, 1980, 17(4), 323-349.

Brzeinski, J., & Schoephoerster, H. Basic reading tests for Images. Boston: Houghton-Mifflin, 1974.

Carver, R. P. Special problems in measuring change with psychometric devices. Evaluation research: Strategies and methods. Pittsburgh: American Institute for Research, 1970.

Deno, S. L., Mirkin, P. K., & Chiang, B. Identifying valid measures of reading. Exceptional Children, 1982, 49(1), 36-45.

Fuchs, L. S., Deno, S. L., & Marston, D. Improving the reliability of curriculum-based measures of academic skills for psychoeducational decision making. Diagnostique, in press.

Fuchs, L. S., Fuchs, D., & Deno, S. L. Reliability and validity of curriculum-based informal reading inventories. Reading Research Quarterly, 1982, 18(1), 6-26.

Fuchs, L. S., Wesson, C., Tindal, G., Mirkin, P. K., & Deno, S. L. Teacher efficiency in continuous evaluation of IEP goals (Research Report No. 53). Minneapolis: University of Minnesota, Institute for Research on Learning Disabilities, 1981.

Glaser, R., & Nitko, J. Measurement in learning and instruction. In R. L. Thorndike (Ed.), Educational measurement (2nd ed.). Washington, D.C.: American Council on Education, 1971.

Goodstein, H. A. The reliability of criterion-referenced tests and special education: Assumed versus demonstrated. Journal of Special Education, 1982, 16 (1), 37-48.

Hambleton, R. K. Test score validity. In R. A. Berk (Ed.), Criterion-referenced measurement: The state of the art. Baltimore: The Johns Hopkins University Press, 1980.

Hambleton, R. K., & Novick, M. R. Toward an integration of theory and method for criterion-referenced tests. Journal of Educational Measurement, 1973, 10(3), 159-170.

Hambleton, R. K., Swaminathan, H., Algina, J., & Coulson, D. B. Criterion-referenced testing and measurement: A review of technical issues and developments. Review of Educational Research, 1978, 48(1), 1-47.

Huynh, H. On the reliability of decisions in domain-referenced testing. Journal of Educational Measurement, 1976, 13, 253-264.

Millman, J. Criterion-referenced measurement. In W. J. Popham (Ed.), Evaluation in education: Current applications. Berkeley: McCutchan, 1974.

Naslund, R. A., Thorpe, L. P., & Lefever, D. W. SRA achievement Series: Reading, mathematics, and language arts. Chicago: Science Research Associates, 1978.

Popham, W. J. Domain specification strategies. In R. A. Berk (Ed.), Criterion-referenced measurement: The state of the art. Baltimore: The Johns Hopkins University Press, 1980.

Popham, W. J., & Husek, T. R. Implications of criterion-referenced measurement. Journal of Educational Measurement, 1969, 6, 1-9.

Salvia, J., & Ysseldyke, J. E. Assessment in special and remedial education (2nd ed.). Boston: Houghton-Mifflin, 1981.

Skager, R. The system for objectives-based evaluation--Reading. Evaluation Comment, 1971, 3, 6-11.

Subkoviak, M. J. Estimating reliability from a single administration of a mastery test. Madison, Wis.: Laboratory of Experimental Design, University of Wisconsin, 1975.

Footnote

[1]In response to a written request for information concerning the technical adequacy of the test studied here, publishers described the field-testing that they had conducted. This response (a) alluded to, but failed to describe, an item analysis of test data, and (b) reported on a pre-posttest study in which students demonstrated an average growth of 8.5 grade equivalent months in 7 chronological months on the Gates-MacGinitie. Authors of the response stated that "This tends to confirm that the use of [criterion-referenced] tests...to monitor effectiveness of instruction and reteaching contributed to an appropriate rate of progress among students."

Table 1

Traditional and Alternative Studies of Reliability and Validity Reported in Manuals of Commercial Criterion-referenced Tests and Basal Series Mastery Tests

Tests	Reported in Test Manuals[a]							
	Traditional (correlational) analyses						Alternative (criterion-referenced) analyses	
	reliability				validity			
	inter-rater	alternate-form	internal consistency	test-retest	construct	criterion	reliability studies	validity studies
Commercial								
Brigance Diagnostic Inventory of Basic Skills (1977)								
Brigance Diagnostic Inventory of Early Development (1977)								
CIRCUS (1974)		X	X		X	X		
Curriculum and Monitoring System								
Developmental Programming for infants and young children: Assessment & Application (1977)				X	X	X		
Kindergarten Evaluation of Learning Potential: A curricular approach to evaluation (1963)						X		
Learning Accomplishment Profile (1977)	X							
Learning Staircase (1976)								
Basal Series Mastery								
Ginn 720 (1979)								
Holt (1981)								
Houghton-Mifflin (1974)								
Scott Foresman (1981)								

[a]X indicates that a study was reported in the test manual.

Table 2

Examinees' Tasks on the Houghton-Mifflin End-of-Level 11 Basic Reading Test

Scale/Subtest	Examinees' Tasks
Decoding	
Word Attack	1. Read a sentence from which letter(s) of one word have been deleted. From an array of three choices, circle the word that most nearly sounds like the unfinished word.
Pronunciation	1. Given a word in dictionary spelling, select from three choices the word(s) with the same vowel sounds as the dictionary-spelled word.
Comprehension	
Literal Comprehension	1. Read a factual article comprising four paragraphs. Then, identify each of 12 statements as either true or false with respect to information provided in the article.
Interpretive Thinking	1. Read a paragraph, and (a) select the main idea from a set of statements, and (b) determine whether each distractor is not the main idea because the paragraph either fails to address the statement or is broader than the statement.
Meaning Acquisition	1. Given a sentence with an underlined word and given meanings for the underlined word, select the meaning that best fits the sentence.
	2. Given a sentence with an underlined figure of speech, select from a set of possible statements the one best defining the figure of speech in the sentence.
	3. Given a sentence with an underlined word containing a common prefix and given three possible meanings, select the best meaning for the underlined word.

Table 2 (continued)

Scale/Subtest	Examinees' Tasks
Reference/Study Skills	
Information Locating	1. Given a book's abbreviated index and a set of questions, write page numbers of the book on which a relevant answer might be located for each question.
	2. Given questions and an illustration of a 21-volume encyclopedia, write the volume number in which relevant information might be located for each question. Then, given questions and a list of possible subheadings for the topic Newspaper, write the subheading in which a relevant answer might be located for each question.
	3. Given questions and an illustration of a card catalog, identify the drawer in which a relevant answer might be located for each question. Then, given questions, determine whether one would search for an author, title, or subject card for a relevant answer to each question.
	4. Given questions and a 5-column, 10-row table containing information on the first 10 presidents, answer each question.
Information Appraising	1. Identify whether statements are fact, fiction, or both.
	2. Given a set of opinion statements and a set of persons with biographical information, match the person best qualified to make each opinion statement.
	3. Identify whether or not a statement contains vague statements, and if so, underline the vague statement.
Information Organizing	1. Read an article. Complete a partially completed outline concerning the article with three levels of information: main topics, subtopics, and details.

Table 3

Student Performance on Measures of Reading Achievement

Test	Mean	SD
End-of-Level 11 Basic Reading Test[a]		
Decoding Subtests		
Word Attack	22.5	3.2
Pronunciation	17.9	6.1
Decoding Composite	40.4	8.0
Comprehension Subtests		
Literal Comprehension	20.2	3.8
Interpretive Thinking	19.7	5.5
Meaning Acquisition	62.3	11.5
Comprehension Composite	102.2	17.8
Study/Reference Skills Subtests		
Information Locating	79.3	17.9
Information Appraising	45.1	17.3
Information Organizing	18.0	8.6
Reference/Study Skill Composite	142.4	38.9
SRA Reading Achievement Test[b]		
Vocabulary	23.4	8.6
Comprehension	28.8	11.1
Total	51.5	18.1
Word Reading Test[c]		
Isolated Word Reading	46.6	18.4
Passage Reading	117.8	34.5

[a] N = 46

[b] N = 42

[c] N = 47

Table 4

Test-retest Reliabilities for Houghton-Mifflin End-of-level 11 Basic Reading Test (N=20)

Subtest	Reliability
Decoding Subtests	
Word Attack	.42
Pronunciation	.20
Decoding Composite	.21
Comprehension Subtests	
Literal Comprehension	.61
Interpretive Thinking	.03
Meaning Acquisition	.83
Comprehension Composite	.72
Study/Reference Skills Subtests	
Information Locating	.94
Information Appraising	.86
Information Organizing	.93
Reference/Study Skill Composite	.94

Table 5

Proportion of Subjects with Varying Percentages of Difference Scores Across Two Administrations of the End-of-level 11 Basic Reading Test (N=20)

Basic Reading Test	N[a]	Percentage Difference Score									
		0 to .07	.08 to .14	.15 to .24	.25 to .34	.35 to .44	.45 to .54	.55 to .64	.65 to .74	.75 to .84	.85 to 1.0
Decoding Subtests											
Word Attack	6	85	0	10	5	0	0	0	0	0	0
Pronunciation	8	45	27	0	15	7	6	0	0	0	0
Comprehension Subtests											
Literal Comprehension	12	38	35	15	12	0	0	0	0	0	0
Interpretive Thinking	12	55	10	10	8	7	0	0	5	5	0
Meaning Acquisition											
Words	12	50	40	5	5	0	0	0	0	0	0
Figures of Speech	12	77	23	0	0	0	0	0	0	0	0
Affixes	12	66	19	5	5	0	5	0	0	0	0
Study/Reference Skills Subtests											
Information Locating											
Index	12	77	15	8	0	0	0	0	0	0	0
Encyclopedia	12	38	50	0	6	6	0	0	0	0	0
Card Catalog	12	50	27	17	6	0	0	0	0	0	0
Table	12	55	25	20	0	0	0	0	0	0	0
Information Appraising											
Fact/Fiction	12	22	56	22	0	0	0	0	0	0	0
Opinion Statements	6	55	0	10	25	0	0	0	5	0	0
Value Expressions	6	38	0	50	0	0	12	0	0	0	0
Information Organizing	12	75	10	5	5	5	0	0	0	0	0

[a]Number of items on the test.

Table 6

Uncorrected and Corrected Proportion of Examinees (N=18) Placed Into the Same Decision Categories on Two Administrations of the End-of-level 11 Basic Reading Test

Basic Reading Test	Proportion of Examinees	
	Uncorrected	Corrected for Chance Agreements[a]
Decoding Subtests		
Word Attack	.89	-.06
Pronunciation	.61	.18
Comprehension Subtests		
Literal Comprehension	.83	.57
Interpretive Thinking	.72	-.15
Meaning Acquisition		
Words	.72	.31
Figures of Speech	.89	.68
Affixes	.94	.88
Study/Reference Skills Subtests		
Information Locating		
Index	.89	.68
Encyclopedia	.89	.72
Card Catalog	.83	.47
Table	.89	.68
Information Appraising		
Fact/Fiction	.89	.68
Opinion Statements	.89	.78
Value Expressions	.78	.51
Information Organizing	.89	.78

[a]Observed - Chance Proportions/Maximum Value that (Observed-Chance Proportions) Can Assume.

Table 7

Correlations Between Basic Reading Test and SRA Test Scores (N=42)

Basic Reading Test	SRA Vocabulary	SRA Comprehension	SRA Total
Decoding Subtests			
Word Attack	.40	.38	.40
Pronunciation	.42	.44	.43
Decoding Composite	.48	.49	.49
Comprehension Subtests			
Literal Comprehension	.52	.61	.57
Interpretive Thinking	.35	.19	.26
Meaning Acquisition	.73	.70	.75
Comprehension Composite	.70	.64	.69
Study/Reference Skills Subtests			
Information Locating	.67	.63	.65
Information Appraising	.58	.55	.53
Information Organizing	.54	.47	.51
Reference/Study Skill Composite	.69	.63	.65

Table 8

Correlations Between Basic Reading Test and Word Reading Test Scores (N=46)

Basic Reading Test Subtests	Word Reading Test	
	Isolated Words	Passage
Decoding Subtests		
Word Attack	.27	.31
Pronunciation	.33	.45
Decoding Composite	.36	.47
Comprehension Subtests		
Literal Comprehension	.41	.50
Interpretive Thinking	.33	.37
Meaning Acquisition	.55	.67
Comprehension Composite	.55	.66
Study/Reference Skills Subtests		
Information Locating	.53	.64
Information Appraising	.48	.59
Information Organizing	.52	.57
Reference/Study Skills Composite	.57	.68
Total Test Score	.57	.65

Table 9

Relation Between Houghton-Mifflin Basic Reading Tests and Criterion Classification (N=46)

Basic Reading Tests	χ^2	p-value	Percentage Misclassified
Decoding Subtests			
Word Attack	2.3	.15	43
Pronunciation	1.8	.22	38
Decoding Composite	5.1	.03	32
Comprehension Subtests			
Literal Comprehension	1.5	.25	40
Interpretive Thinking	.8	.40	43
Meaning Acquisition	4.6	.04	34
Comprehension Composite	5.1	.03	32
Study/Reference Skills Subtests			
Information Locating	5.4	.02	32
Information Acquiring	20.7	<.001	15
Information Organizing	11.5	<.001	23
Reference/Study Skills Composite	11.7	<.001	23

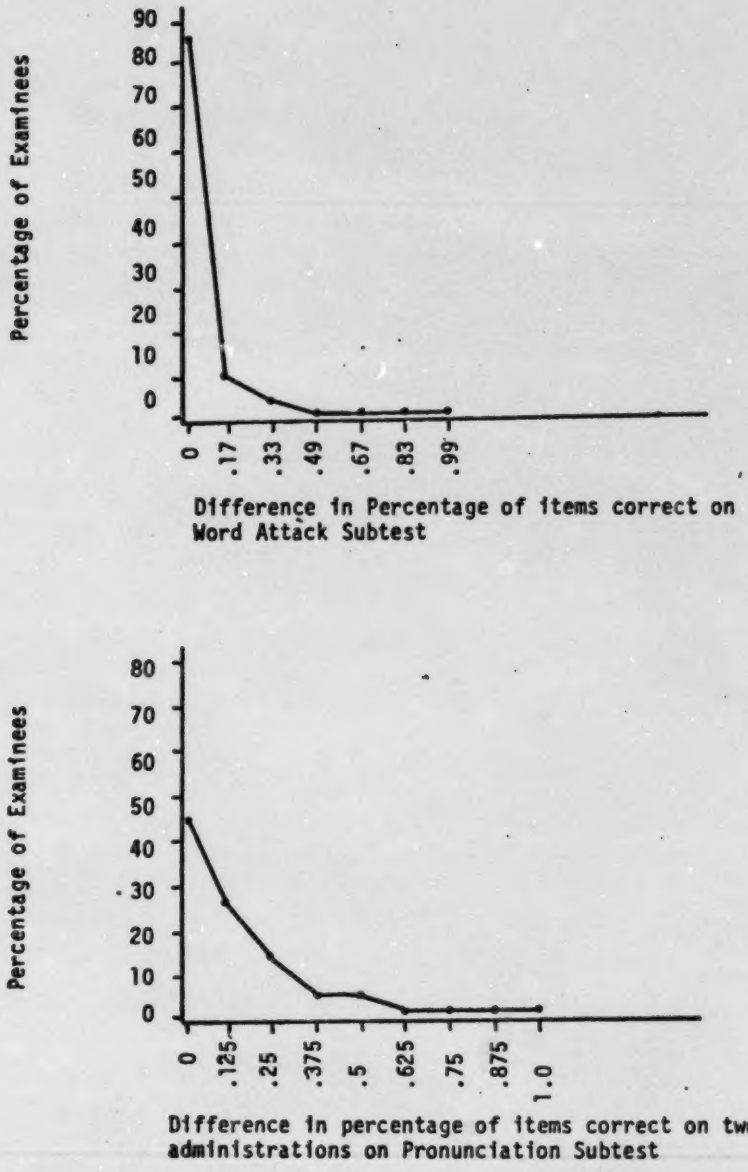


Figure 1. Displays of consistency of test scores on decoding subtests of end-of-level 11 BRT.

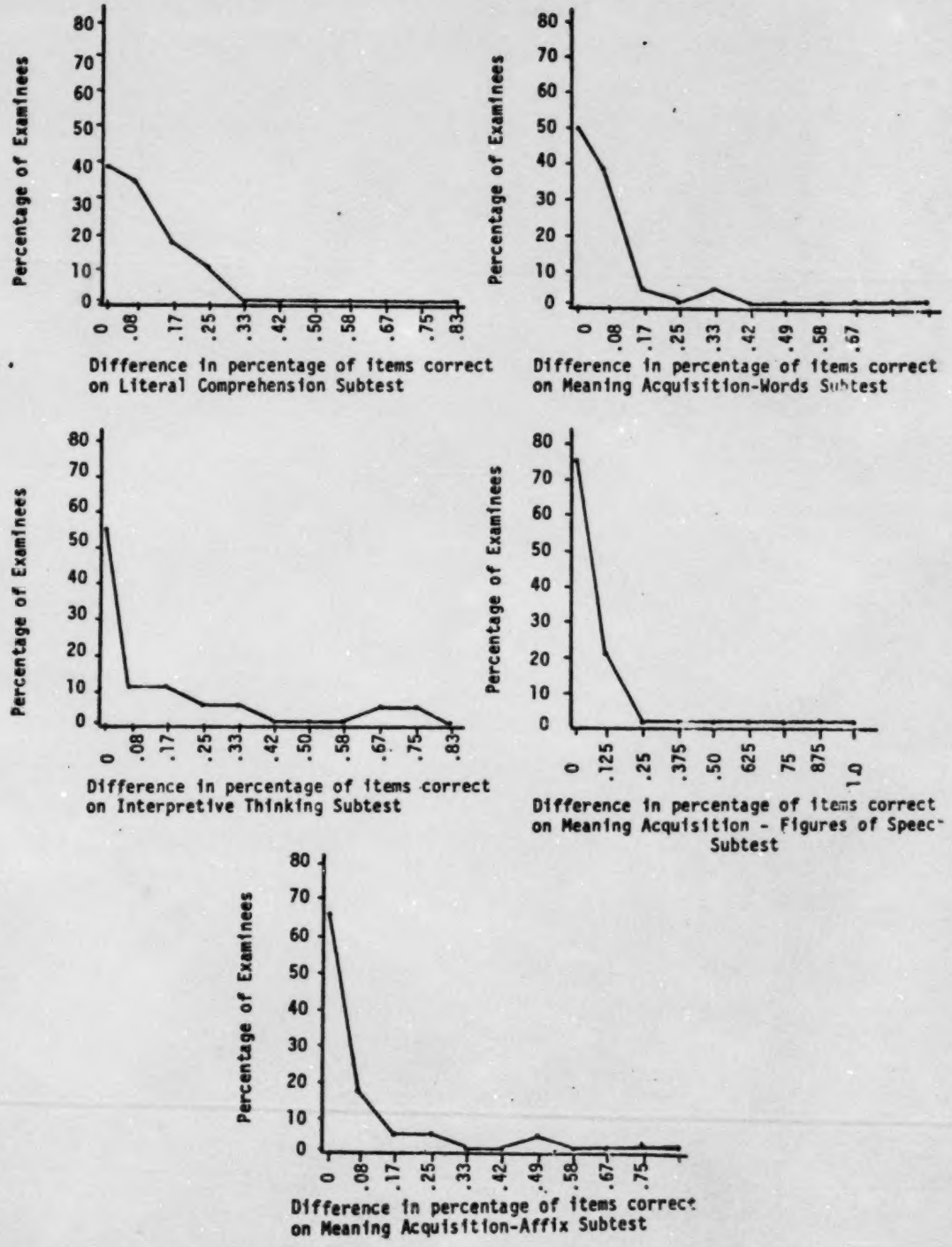
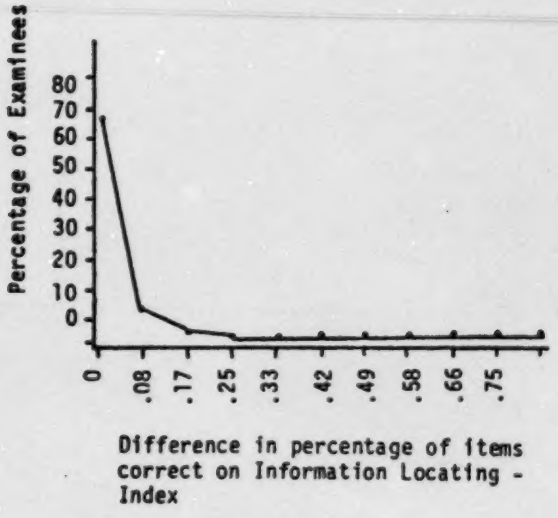
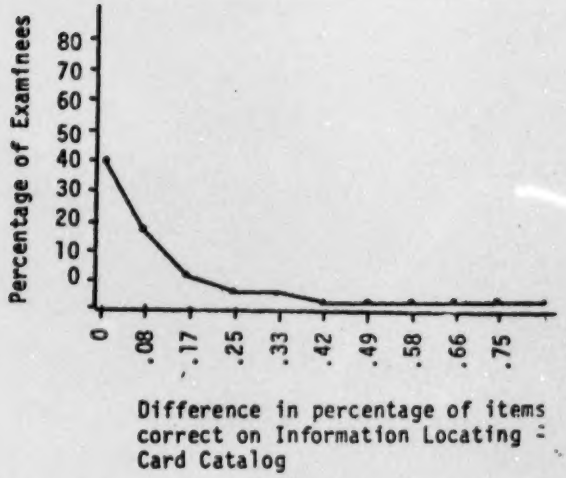
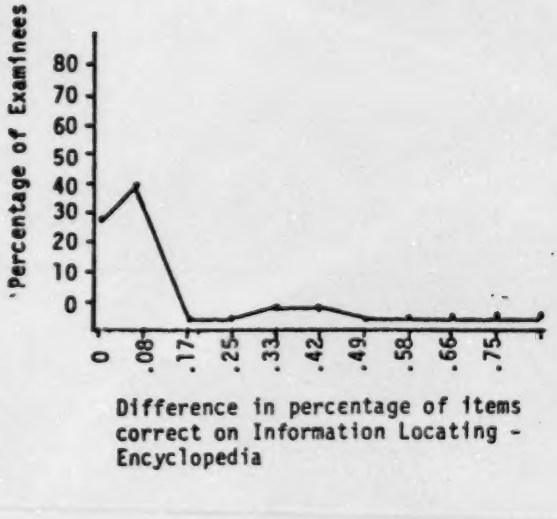

Percentage of Examinees
80
70
60
50
40
30
20
10
0
0 .08 .17 .25 .33 .42 .50 .58 .67 .75 .83
Difference in percentage of items correct on Literal Comprehension Subtest
Percentage of Examinees
0 .08 .17 .25 .33 .42 .49 .58 .67
Difference in percentage of items correct on Meaning Acquisition-Words Subtest
Percentage of Examinees
0 .08 .17 .25 .33 .42 .50 .58 .67 .75 .83
Difference in percentage of items correct on Interpretive Thinking Subtest
Percentage of Examinees
0 .125 .25 .375 .50 .625 .75 .875 1.0
Difference in percentage of items correct on Meaning Acquisition - Figures of Speech Subtest
Percentage of Examinees
0 .08 .17 .25 .33 .42 .49 .58 .67 .75
Difference in percentage of items correct on Meaning Acquisition-Affix Subtest

Figure 2. Displays of consistency of test scores on comprehension subtests of end-of-level 11 BRT.







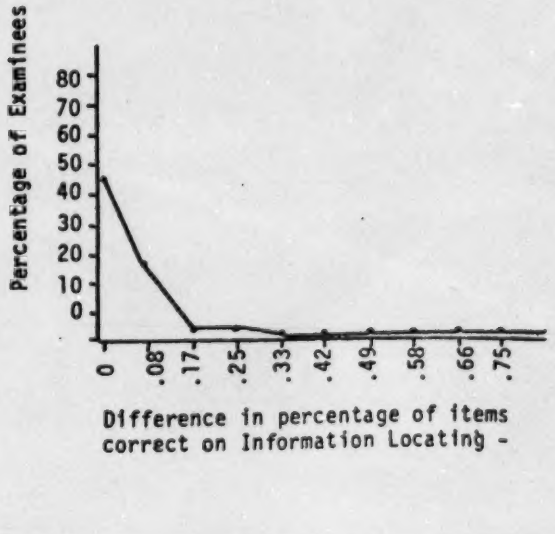
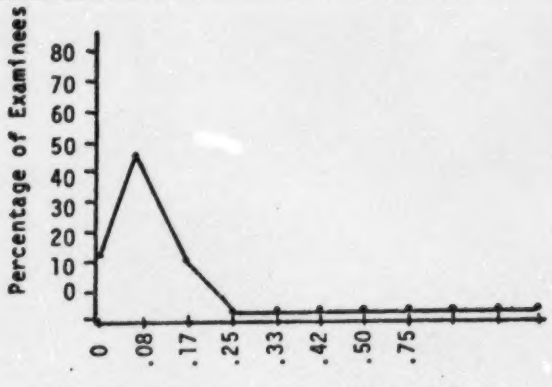
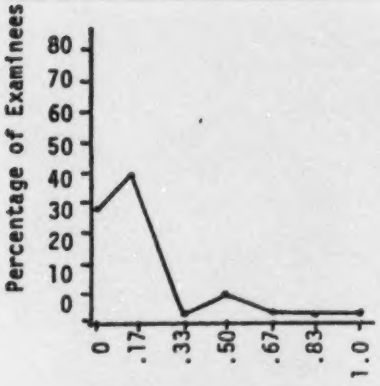


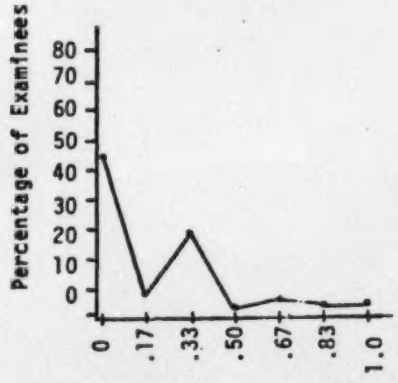
Figure 3. Displays of consistency of test scores on study/reference skills, information locating subtests of end-of-level 11 BRT.



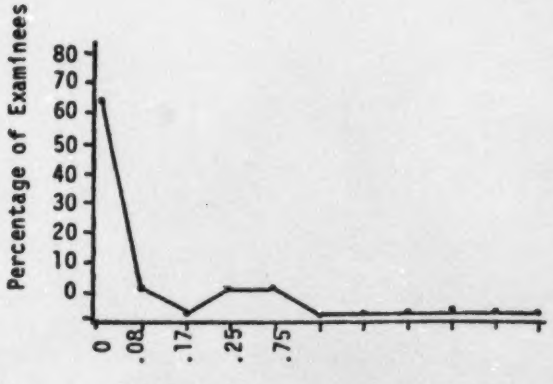
Difference in percentage of items correct on Information Appraising-Fact/Fiction



Difference in percentage of items correct on Information Appraising - Vague Expressions



Difference in percentage of items correct on Information Appraising - Evaluation Statements



Difference in percentage of items correct on Information Organizing

Figure 4. Displays of consistency of test scores on study/reference skills, information appraising and information organizing subtests of end-of-level 11 BRT.

PUBLICATIONS

Institute for Research on Learning Disabilities
University of Minnesota

The Institute is not funded for the distribution of its publications. Publications may be obtained for $4.00 each, a fee designed to cover printing and postage costs. Only checks and money orders payable to the University of Minnesota can be accepted. All orders must be prepaid. Requests should be directed to: Editor, IRLD, 350 Elliott Hall; 75 East River Road, University of Minnesota, Minneapolis, MN 55455.

The publications listed here are only those that have been prepared since 1982. For a complete, annotated list of all IRLD publications, write to the Editor.

Wesson, C., Mirkin, P., & Deno, S. Teachers' use of self instructional materials for learning procedures for developing and monitoring progress on IEP goals (Research Report No. 63). January, 1982.

Fuchs, L., Wesson, C., Tindal, G., Mirkin, P., & Deno, S. Instructional changes, student performance, and teacher preferences: The effects of specific measurement and evaluation procedures (Research Report No. 64). January, 1982.

Potter, M., & Mirkin, P. Instructional planning and implementation practices of elementary and secondary resource room teachers: Is there a difference? (Research Report No. 65). January, 1982.

Thurlow, M. L., & Ysseldyke, J. E. Teachers' beliefs about LD students (Research Report No. 66). January, 1982.

Graden, J., Thurlow, M. L., & Ysseldyke, J. E. Academic engaged time and its relationship to learning: A review of the literature (Monograph No. 17). January, 1982.

King, R., Wesson, C., & Deno, S. Direct and frequent measurement of student performance: Does it take too much time? (Research Report No. 67). February, 1982.

Greener, J. W., & Thurlow, M. L. Teacher opinions about professional education training programs (Research Report No. 68). March, 1982.

Algozzine, B., & Ysseldyke, J. Learning disabilities as a subset of school failure: The oversophistication of a concept (Research Report No. 69). March, 1982.

Fuchs, D., Zern, D. S., & Fuchs, L. S. A microanalysis of participant behavior in familiar and unfamiliar test conditions (Research Report No. 70). March, 1982.

Shinn, M. R., Ysseldyke, J., Deno, S., & Tindal, G. A comparison of psychometric and functional differences between students labeled learning disabled and low achieving (Research Report No. 71). March, 1982.

Thurlow, M. L. Graden, J., Greener, J. W., & Ysseldyke, J. E. Academic responding time for LD and non-LD students (Research Report No. 72). April, 1982.

Graden, J., Thurlow, M., & Ysseldyke, J. Instructional ecology and academic responding time for students at three levels of teacher-perceived behavioral competence (Research Report No. 73). April, 1982.

Algozzine, B., Ysseldyke, J., & Christenson, S. The influence of teachers' tolerances for specific kinds of behaviors on their ratings of a third grade student (Research Report No. 74). April, 1982.

Wesson, C., Deno, S., & Mirkin, P. Research on developing and monitoring progress on IEP goals: Current findings and implications for practice (Monograph No. 18). April, 1982.

Mirkin, P., Marston, D., & Deno, S. L. Direct and repeated measurement of academic skills: An alternative to traditional screening, referral, and identification of learning disabled students (Research Report No. 75). May, 1982.

Algozzine, B., Ysseldyke, J., Christenson, S., & Thurlow, M. Teachers' intervention choices for children exhibiting different behaviors in school (Research Report No. 76). June, 1982.

Tucker, J., Stevens, L. J., & Ysseldyke, J. E. Learning disabilities: The experts speak out (Research Report No. 77). June, 1982.

Thurlow, M. L., Ysseldyke, J. E., Graden, J., Greener, J. W., & Mecklenberg, C. Academic responding time for LD students receiving different levels of special education services (Research Report No. 78). June, 1982.

Graden, J. L., Thurlow, M. L., Ysseldyke, J. E., & Algozzine, B. Instructional ecology and academic responding time for students in different reading groups (Research Report No. 79). July, 1982.

Mirkin, P. K., & Potter, M. L. A survey of program planning and implementation practices of LD teachers (Research Report No. 80). July, 1982.

Fuchs, L. S., Fuchs, D., & Warren, L. M. Special education practice in evaluating student progress toward goals (Research Report No. 81). July, 1982.

Kuehnle, K., Deno, S. L., & Mirkin, P. K. Behavioral measurement of social adjustment: What behaviors? What setting? (Research Report No. 82). July, 1982.

Fuchs, D., Dailey, Ann Madsen, & Fuchs, L. S. Examiner familiarity and the relation between qualitative and quantitative indices of expressive language (Research Report No. 83). July, 1982.

Videen, J., Deno, S., & Marston, D. Correct word sequences: A valid indicator of proficiency in written expression (Research Report No. 84). July, 1982.

Potter, M. L. Application of a decision theory model to eligibility and classification decisions in special education (Research Report No. 85). July, 1982.

Greener, J. E., Thurlow, M. L., Graden, J. L., & Ysseldyke, J. E. The educational environment and students' responding times as a function of students' teacher-perceived academic competence (Research Report No. 86). August, 1982.

Deno, S., Marston, D., Mirkin, P., Lowry, L., Sindelar, P., & Jenkins, J. The use of standard tasks to measure achievement in reading, spelling, and written expression: A normative and developmental study (Research Report No. 87). August, 1982.

Skiba, R., Wesson, C., & Deno, S. L. The effects of training teachers in the use of formative evaluation in reading: An experimental-control comparison (Research Report No. 88). September, 1982.

Marston, D., Tindal, G., & Deno, S. L. Eligibility for learning disability services: A direct and repeated measurement approach (Research Report No. 89). September, 1982.

Thurlow, M. L., Ysseldyke, J. E., & Graden, J. L. LD students' active academic responding in regular and resource classrooms (Research Report No. 90). September, 1982.

Ysseldyke, J. E., Christenson, S., Pianta, R., Thurlow, M. L., & Algozzine, B. An analysis of current practice in referring students for psychoeducational evaluation: Implications for change (Research Report No. 91). October, 1982.

Ysseldyke, J. E., Algozzine, B., & Epps, S. A logical and empirical analysis of current practices in classifying students as handicapped (Research Report No. 92). October, 1982.

Tindal, G., Marston, D., Deno, S. L., & Germann, G. Curriculum differences in direct repeated measures of reading (Research Report No. 93). October, 1982.

Fuchs, L.S., Deno, S. L., & Marston, D. Use of aggregation to improve the reliability of simple direct measures of academic performance (Research Report No. 94). October, 1982.

Ysseldyke, J. E., Thurlow, M. L., Mecklenburg, C., & Graden, J. Observed changes in instruction and student responding as a function of referral and special education placement (Research Report No. 95). October, 1982.

Fuchs, L. S., Deno, S. L., & Mirkin, P. K. Effects of frequent curriculum-based measurement and evaluation on student achievement and knowledge of performance: An experimental study (Research Report No. 96). November, 1982.

Fuchs, L. S., Deno, S. L., & Mirkin, P. K. Direct and frequent measurement and evaluation: Effects on instruction and estimates of student progress (Research Report No. 97). November, 1982.

Tindal, G., Wesson, C., Germann, G., Deno, S. L., & Mirkin, P. K. The Pine County model for special education delivery: A data-based system (Monograph No. 19), November, 1982.

Epps, S., Ysseldyke, J. E., & Algozzine, B. An analysis of the conceptual framework underlying definitions of learning disabilities (Research Report No. 98). November, 1982.

Epps, S., Ysseldyke, J. E., & Algozzine, B. Public-policy implications of different definitions of learning disabilities (Research Report No. 99). November, 1982.

Ysseldyke, J. E., Thurlow, M. L., Graden, J. L., Wesson, C., Deno, S. L., & Algozzine, B. Generalizations from five years of research on assessment and decision making (Research Report No. 100). November, 1982.

Marston, D., & Deno, S. L. Measuring academic progress of students with learning difficulties: A comparison of the semi-logarithmic chart and equal interval graph paper (Research Report No. 101). November, 1982.

Beattie, S., Grise, P., & Algozzine, B. Effects of test modifications on minimum competency test performance of third grade learning disabled students (Research Report No. 102). December, 1982

Algozzine, B., Ysseldyke, J. E., & Christenson, S. An analysis of the incidence of special class placement: The masses are burgeoning (Research Report No. 103). December, 1982.

Marston, D., Tindal, G., & Deno, S. L. Predictive efficiency of direct, repeated measurement: An analysis of cost and accuracy in classification (Research Report No. 104). December, 1982.

Wesson, C., Deno, S., Mirkin, P., Sevcik, B., Skiba, R., King, R., Tindal, G., & Maruyama, G. Teaching structure and student achievement effects of curriculum-based measurement: A causal (structural) analysis (Research Report No. 105). December, 1982.

Mirkin, P. K., Fuchs, L. S., & Deno, S. L. (Eds.). Considerations for designing a continuous evaluation system: An integrative review (Monograph No. 20). December, 1982.

Marston, D., & Deno, S. L. Implementation of direct and repeated measurement in the school setting (Research Report No. 106). December, 1982.

Deno, S. L., King, R., Skiba, R., Sevcik, B., & Wesson, C. The structure of instruction rating scale (SIRS): Development and technical characteristics (Research Report No. 107). January, 1983.

Thurlow, M. L., Ysseldyke, J. E., & Casey, A. Criteria for identifying LD students: Definitional problems exemplified (Research Report No. 108). January, 1983.

Tindal, G., Marston, D., & Deno, S. L. The reliability of direct and repeated measurement (Research Report No. 108). February, 1983.

Fuchs, D., Fuchs, L. S., Dailey, A. M., & Power, M. H. Effects of pre-test contact with experienced and inexperienced examiners on handicapped children's performance (Research Report No. 110). February, 1983

King, R. P., Deno, S., Mirkin, P., & Wesson, C. The effects of training teachers in the use of formative evaluation in reading: An experimental-control comparison (Research Report No. 111). February, 1983.

Tindal, G., Deno, S. L., & Ysseldyke, J. E. Visual analysis of time series data: Factors of influence and level of reliability (Research Report No. 112). March, 1983.

Tindal, G, Shinn, M., Fuchs, L., Fuchs, D., Deno, S., & Germann, G. The technical adequacy of a basal reading series mastery test (Research Report No. 113). April, 1983.